PAN ARCHIVES OF ACOUSTICS
Vol. **42**, No. 3, pp. 365–373 (2017)

DOI: 10.1515/aoa-2017-0038

Head-Related Transfer Function Selection Using Neural Networks

Shu-Nung YAO[(1)], Tim COLLINS[(2)], Chaoyun LIANG[(3)]

[(1)] *Department of Electrical Engineering*
National Taipei University
No. 151, University Rd., San Shia District, New Taipei City 23741, Taiwan; e-mail: snyao@gm.ntpu.edu.tw

[(2)] *School of Engineering*
Manchester Metropolitan University
Manchester, M1 5GD, UK; e-mail: T.Collins@mmu.ac.uk

[(3)] *Department of Bio-Industry Communication and Development*
National Taiwan University
No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan; e-mail: cliang@ntu.edu.tw

(*received August 1, 2016; accepted April 4, 2017*)

In binaural audio systems, for an optimal virtual acoustic space a set of head-related transfer functions (HRTFs) should be used that closely matches the listener's ones. This study aims to select the most appropriate HRTF dataset from a large database for users without the need for extensive listening tests. Currently, there is no way to reliably reduce the number of datasets to a smaller, more manageable number without risking discarding potentially good matches. A neural network that estimates the appropriateness of HRTF datasets based on input vectors of anthropometric measurements is proposed. The shapes and sizes of listeners' heads and pinnas were measured using digital photography; the measured anthropometric parameters form the feature vectors used by the neural network. A graphical user interface (GUI) was developed for participants to listen to music transformed using different HRTFs and to evaluate the fitness of each HRTF dataset. The listening scores recorded were the target outputs used to train the neural networks. The aim was to learn a mapping between anthropometric parameters and listener's perception scores. Experimental validations were performed on 30 subjects. It is demonstrated that the proposed system produces a much more reliable HRTF selection than previously used methods.

Keywords: head-related transfer function; neural networks; localisation; music; audio; anthropometry; pinna.

1. Introduction

Head-related transfer function (HRTF) datasets are different for different people; efficiently obtaining an appropriate HRTF dataset for a user is an active area of research. The potential problems of using non-individualised HRTFs were discussed in Yao and Chen (2013). If the HRTF dataset in a binaural audio device does not match the real one of the user, poor localisation happens. The direct way to obtain individualised HRTFs is to perform HRTF measurements (Gardner, Martin, 1994; Algazi *et al.*, 2001), but this is expensive and time consuming. Another approach is to mathematically model the acoustic properties of the head and ear. The spectral cues are caused by the reflection and absorption from parts of the listener's body, such as the concha, head, and torso. Several mathematical models were developed to model how sound waves are affected by the geometry of a human body. Brown and Duda (1998) used Rayleigh's spherical model to generate the interaural time delay (ITD) cues and a single-pole, single-zero head-shadow filter to produce the interaural level difference (ILD) cues. Batteau (1967) pointed out the direct sound always accompanies multiple echoes when the sound source travels to arrive at the ear canal. This is because of the delays caused by the reflecting surfaces of pinna. Watkins (1978) formalised Batteau's pinna model (Batteau, 1967) as shown in Fig. 1. ρ_A and ρ_V are reflection coefficients and τ_A and τ_B are time delays. The complex geometry of the pinna and inner ear is difficult to model accurately. Watkins (1978)

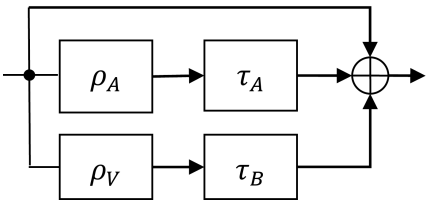

Fig. 1. Batteau's pinna model: ρ – reflection coefficient, τ – time delay.

set ρ_A and ρ_V to unity. τ_A depends on the azimuth angle and is between 0 and 80 μs. τ_B depends on the elevation angle and is between 100 and 320 µs. Brown and Duda (1997) empirically designed a pinna model by the examination of the head-related impulse response (HRIR) database. Their structure is presented in Fig. 2. There are five reflection paths. Typical values of the gains ρ_k are shown in Table 1 and the delays $\tau_k(\theta, \varphi)$ are in the form of

$$\tau_k(\theta, \varphi) = A_k \cos\left(\frac{\theta}{2}\right) \sin\left(D_k\left(\frac{\pi}{2} - \varphi\right)\right) + B_k, \quad (1)$$

where θ stands for azimuthal angle and φ, altitude angle. A_k, B_k, and D_k are constants; typical values can be found in Table 1.

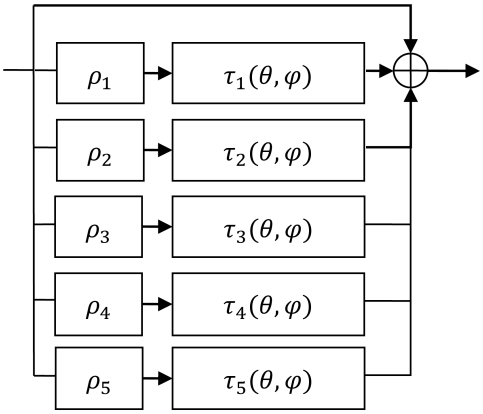

Fig. 2. Modern pinna model: ρ – reflection coefficient, τ – time delay, θ – azimuth, φ – elevation.

Table 1. Example values of the pinna model parameters.

k	ρ_k	A_k	B_k	D_k
1	0.50	1	2	1.0
2	−1.00	5	4	0.5
3	0.50	5	7	0.5
4	−0.25	5	11	0.5
5	0.25	5	13	0.5

Another approach to HRTF personalisation is to select the closest matching HRTF dataset from an existing database. There are several HRTF databases (Algazi *et al.*, 2001; Ircam, 2002; Gupta *et al.*, 2010) available online, some of which release their participants' anthropometric parameters such as head sizes and ear sizes. HRTF dataset selection based on seven anthropometric parameters of the pinna was proposed in Zotkin *et al.* (2004). The seven anthropometric parameters are shown in Fig. 3. A distance measure based on the sum of the squared errors between the measurements of the listener and each of the database members was calculated as

$$E_h = \sum_{i=1}^{7} \left(\frac{\widehat{d}(i) - d_h(i)}{D(i)}\right)^2, \quad (2)$$

where $\widehat{d}(i)$ and $d_h(i)$ are the i-th anthropometric parameters of the listener and of the h-th member of the database, and $D(i)$ is the standard deviation of the i-th anthropometric parameter, estimated over all the database members. The HRTF dataset corresponding to the minimum E_h was selected for the listener. The method was developed on the basis of the idea that each anthropometric parameter is equally important to listening perception. However, the influence of anthropometric parameters is complicated, and HRTF spectrum might be more sensitive to certain parameters. As a result, we chose multi-layer perceptron to deal with non-linear and multivariate functions.

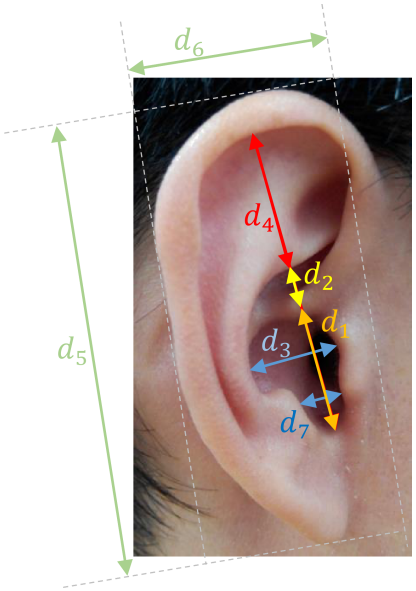

Fig. 3. Pinna model and its parameters: d_1 – cavum concha height, d_2 – cymba concha height, d_3 – cavum concha width, d_4 – fossa height, d_5 – pinna height, d_6 – pinna width, d_7 – intertragal incisure width.

2. Experimental method

This paper aims to find a more reliable method of selecting an appropriate HRTF dataset from an existing database. First, we randomly chose 18 HRTF datasets together with their anthropometric parameters from the CIPIC database (Algazi *et al.*, 2001). Secondly, subjects were asked to rate each of the 18

HRTF datasets through a listening test. After the listening test, we measured the size of subjects' heads and took photos for their pinnas, so both their listening scores and personal anthropometric parameters (Fig. 3) were recorded. From the results of these experiments, a neural network was trained to predict the level of HRTF fitness using just the anthropometric parameters alone.

Listening tests were conducted to generate the training data for the neural network. The listening tests focused on front-back and up-down discriminations. In each case, the sounds to be distinguished lay on the same cone of confusion and accurate HRTFs are especially required for discrimination (Yao, Chen, 2012). In the first part of the experiment, we fixed the azimuthal angle and changed the elevation to test up-down discrimination. Then in the second part, the elevation was fixed and the different azimuth angles were presented to test front-back discrimination. Subjects were asked whether or not they could distinguish the two source positions in each test. If the localisation was poor, this indicated that the HRTF dataset was not a good match for the listener.

Although Table 2 shows that 20 is the reasonable number of subjects participating in the listening test, most researchers suggest that the sample size for using a statistical test should be larger than 30 (Pett, 1997; Salkind, 2004). We therefore recruited 18 male subjects and 12 female participants for the experiment. The authors of this paper did not double as subjects. All participants were provided with participant information sheets and signed consent forms in accordance with the ethical conditions overseen by the Science, Technology, Engineering and Mathematics Ethical Review Committee (ERN_13-0124) at the University of Birmingham and Research Ethics Committee (201505HS090) at National Taiwan University.

Table 2. Numbers of participants addressed in the literature survey associate with HRTFs.

References	The number of listeners
Chun *et al.* (2011)	8
Tan and Gan (1998)	10
Choi *et al.* (2011)	12
Gupta *et al.* (2002)	15
Zhang *et al.* (1998)	15
Ranjan and Gan (2015)	18
Shabtai and Rafaely (2014)	19
Masterson *et al.* (2012)	20

In the subjective test, the participants listened to audio files generated using the 18 sets of HRTFs via headphones. A subset of 18 HRTF datasets was selected, rather than using the entire CIPIC database, in order to prevent listeners' fatigue. Each HRTF dataset was used to produce two music files corresponding to each of the two criteria: up-down discrimination and front-back discrimination. There were 36 audio files in total. In the first question, there were two separate sound sources in the median plane with different elevations, as shown in Fig. 4a. Monophonic sound was convolved with the HRIRs corresponding first to the position of the black dot, and then the grey dot. The monophonic sound in our experiment was wide bandwidth piano music. After listening to the two sound sources, listeners were asked how well they could discriminate the sources at the low elevation (the position of the black dot) and then the high elevation (the position of the grey dot). If the source at the high and low elevation can be well discriminated, a high listening score is given. In the second question, listeners were asked to assess the HRTFs from the database in terms of the front-back confusion. The sound sources were placed in the front hemisphere (the position of the black dot) and then symmetrically in the back hemisphere (the position of the grey dot) as indicated in Fig. 4b. Listeners were asked how well they could discriminate the sources coming from the front and then the rear. If the source from the front and the rear can be well discriminated, a high listening score is given. The level of discrimination was recorded using the grade scale as shown in Fig. 5. Each HRTF has two listening scores and the mean was calculated. The user interface for the experiments is shown in Fig. 6.

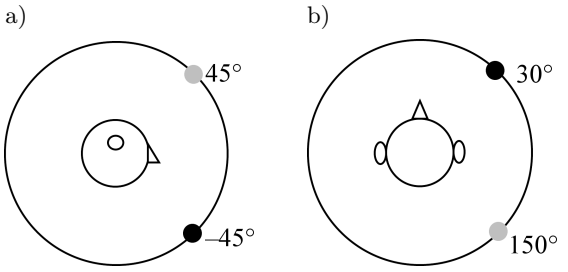

Fig. 4. Experimental setup for: a) up-down confusion, b) front-back confusion.

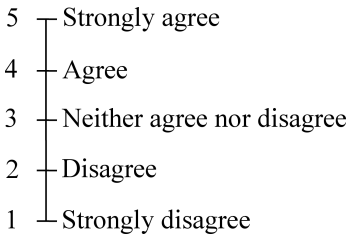

Fig. 5. Five-grade scale for localisation rating: "I can discriminate the source at the high and low elevation" and "I can discriminate the source in front from the source in the back".

Fig. 6. Screen grab of the user interface in the first listening test.

The average listening scores are referred to as the target data and are used to analyse the relationship between anthropometry and individualised HRTFs. The anthropometry contains ten anthropometric parameters, three of which are head width, head height, and head depth as shown in Fig. 7. The remaining seven are cavum concha height, cymba concha height, cavum concha width, fossa height, pinna height, pinna width, and intertragal incisure width from the pinna model in Algazi *et al.*, 2001, also as shown in Fig. 3.

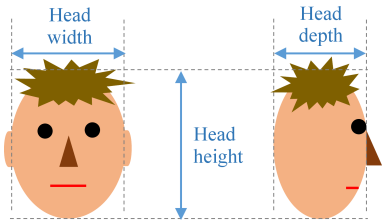

Fig. 7. Anthropometric parameters of head model.

3. Neural network for HRTF selection

The selection of HRTF datasets based on the sum of the squared errors, as presented in Eq. (2), uses a linear contribution from each anthropometric parameter. This may not reflect appropriately the relationship between the person's anthropometric parameters and the most suitable HRTF dataset. We employ neural networks which are capable of modelling complex non-linear relationships. Through training, as shown in Fig. 8, a neural network can learn the best mapping between the users' anthropometric parameters and their listening scores from the given set of training examples and, therefore, predict how suitable each HRTF dataset will be for a new user. Multi-layer neural networks are commonly trained using backpropagation: a gradient decent algorithm, calculating the gradient of a cost function with respect to the weights and the biases in a neural network. They have been broadly used in the vast majority of applications, such as pattern association, pattern classification, data compression, and function approximation (Wythoff, 1993). Two multi-layer feed-forward network structures were created for each HRTF dataset in this paper, the aim being to evaluate which structure provides the better estimate of goodness of fit of each HRTF dataset for a subject.

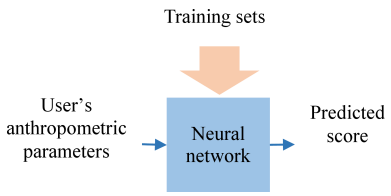

Fig. 8. Neural network training schemes for HRTF selection.

A neural network may contain one or more hidden layers. The most suitable network structure and the number of neurons in each layer are usually found experimentally. However, the network structure should be selected in a way that the total number of parameters in the network is smaller than the number of training data points (Hagan *et al.*, 2002). We experimented with two neural network architectures: a double-hidden-layer neural network and a single-hidden-layer neural network. The double-hidden-layer network, as depicted in Fig. 9, has a total of 21 adjustable parameters, 16 weights and 5 biases. The first hidden layer and the second hidden layer contain one neuron and three neurons, respectively. The single-hidden-layer network, shown in Fig. 10, has a total

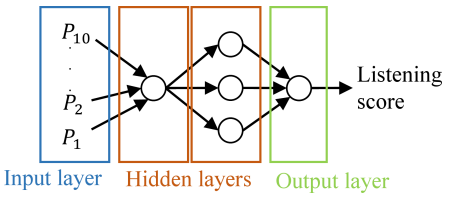

Fig. 9. Double-hidden-layer neural network structure used for HRTF selection.

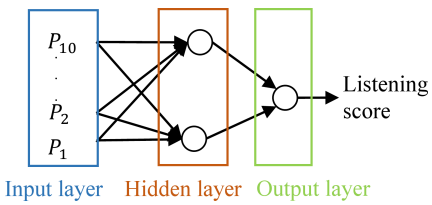

Fig. 10. Single-hidden-layer neural network structure used for HRTF selection.

of 25 adjustable parameters, 22 weights and 3 biases. The hidden layer possesses two neurons. The hyperbolic tangent sigmoid transfer function is used in all hidden layers. There is one neuron in the output layer and this uses the log-sigmoid transfer function to provide a real value in the range (0,1) at the output of the network. This is then scaled and offset to become in the range from 1 to 5 in order to represent the predicted listening score.

We created the neural networks as depicted in Fig. 9 and Fig. 10 for each of the 18 HRTF datasets we had. Each neural network was trained separately. During the training, the network automatically adjusts its parameters, weights and biases, according to the training examples consisting of a set of anthropometric parameters as input features and a listening score as the target output. We applied a leave-one-out procedure to create the feature/target variables. That is, one subject's data was selected as a sample under test, and the others were applied to train the network system. The input features of the training data are relative to participants' anthropometric parameters:

$$\begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,29} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,29} \\ \vdots & \vdots & \vdots & \vdots \\ P_{10,1} & P_{10,2} & \cdots & P_{10,29} \end{bmatrix}, \tag{3}$$

where each column vector $[\, P_{1,k} \;\; P_{2,k} \;\cdots\; P_{10,k} \,]^{\mathrm{T}}$ denotes the ten scaled anthropometric parameters of our k-th participant. The feature scaling is presented in Eq. (4), where $\widehat{d_k}(n)$ is the n-th parameter from the k-th participant' anthropometry, $d_h(n)$ is the n-th anthropometric parameter from the CIPIC's h-th HRTF, and $D(n)$ is the standard deviation of $d_1(n)$, $d_2(n)$, $\ldots$ $d_{18}(n)$, all the n-th anthropometric parameters in the CIPIC database. There are 29 column vectors, because of the leave-one-out procedure. The whole matrix is a requirement to train a neural network.

$$P_{n,k} = \frac{\widehat{d_k}(n) - d_h(n)}{D(n)}, \quad n = 1,\ 2,\ \ldots,\ 10; \quad h = 1,\ 2,\ \ldots,\ 18; \quad k = 1,\ 2,\ \ldots,\ 29. \tag{4}$$

The target variables for the h-th HRTF network, denoted by

$$[\, s_{h,1} \;\; s_{h,2} \;\cdots\; s_{h,29} \,] \tag{5}$$

correspond to the listening scores from our subjects when using the h-th HRTF dataset. The raw listening scores were balanced before used. This is because the same subjective ratings from different listeners can present different meanings. For example, when looking into Subject 2's and Subject 3' scores in Table 3, we found the lowest score is 3. However, when looking into Subject 23's scores, 3 is the highest score. The meaning of 3s in Subject 2's and Subject 3' scores are very different from those in Subject 23's scores. In order to deal fairly with each subject's scores, Eq. (6) is used to balance the listening scores. $S_{\min}$ is the minimum of a subject's scores and $S_{\max}$ is the maximum. After balance, the lowest scores are always equal to 1 and the highest scores are always equal to 5. The balanced scores are shown in Table 4.

$$S_{\mathrm{balanced}}(h) = 1 + 4 \cdot \left(\frac{S(h) - S_{\min}}{S_{\max} - S_{\min}} \right), \quad h = 1,\ 2,\ \ldots,\ 18. \tag{6}$$

During testing, the anthropometric parameters of a new user (who was not included in the training data) were inputted to each of the networks. According to these parameters, the listening score for this user was then predicted by each network giving 18 predicted scores. HRTF datasets were then ranked according to those output scores and the prediction was compared with the participant's actual recorded ranking.

Table 3. Raw data of listening scores.

	HRTF 1	HRTF 2	HRTF 3	HRTF 4	HRTF 5	HRTF 6	HRTF 7	HRTF 8	HRTF 9	HRTF 10	HRTF 11	HRTF 12	HRTF 13	HRTF 14	HRTF 15	HRTF 16	HRTF 17	HRTF 18
Subject 1	4	3.5	4	4	4.5	4.5	4.5	4	2	3.5	4	3.5	3.5	3.5	3	5	4.5	4.5
Subject 2	3.5	3.5	4	3.5	3.5	4	4.5	3.5	4.5	3.5	4	3	3.5	4	4	4	5	4.5
Subject 3	3.5	3	4	3	3	3.5	4	4	3.5	4	3	3.5	3.5	3.5	4.5	3.5	3.5	4.5
Subject 4	3.5	3	3.5	4.5	3	4	4	4	3.5	3	3	4	3.5	3	4	5	4	4
Subject 5	4	3	3.5	3.5	5	4.5	4	4.5	4	3	2.5	3.5	1	3	2.5	2.5	3.5	3
Subject 6	3	3.5	2	3	3	2.5	2.5	2.5	2.5	2	2.5	2.5	3	2	3	2	3	2
Subject 7	3	3.5	3.5	3	3.5	3.5	3.5	4	3	3.5	3	3.5	3.5	3	3	4	3.5	3.5
...																		
Subject 22	2.25	3	2.5	2.5	4	4	4	2.75	4	2.25	3.25	3.75	3.5	2.75	2.5	2.5	2.75	3
Subject 23	1	1.5	1	1	1	1	1.5	1	1	1	3	2.5	2	1	2	2.5	3	2
Subject 24	2.5	3	2.75	2.75	3.25	2.5	2.5	3	2.5	2.75	3	2	1.75	3	2.5	3	2	2.5

Table 4. Balanced listening scores.

	HRTF 1	HRTF 2	HRTF 3	HRTF 4	HRTF 5	HRTF 6	HRTF 7	HRTF 8	HRTF 9	HRTF 10	HRTF 11	HRTF 12	HRTF 13	HRTF 14	HRTF 15	HRTF 16	HRTF 17	HRTF 18
Subject 1	3.67	3	3.67	3.67	4.33	4.33	4.33	3.67	1	3	3.67	3	3	3	2.33	5	4.33	4.33
Subject 2	2	2	3	2	2	3	4	2	4	2	3	1	2	3	3	3	5	4
Subject 3	2.33	1	3.67	1	1	2.33	3.67	3.67	2.33	3.67	1	2.33	2.33	2.33	5	2.33	2.33	5
Subject 4	2	1	2	4	1	3	3	3	2	1	1	3	2	1	3	5	3	3
Subject 5	4	3	3.5	3.5	5	4.5	4	4.5	4	3	2.5	3.5	1	3	2.5	2.5	3.5	3
Subject 6	3.67	5	1	3.67	3.67	2.33	2.33	2.33	2.33	1	2.33	2.33	3.67	1	3.67	1	3.67	1
Subject 7	1	3	3	1	3	3	3	5	1	3	1	3	3	1	1	5	3	3
...																		
Subject 22	1	2.71	1.57	1.57	5	5	5	2.14	5	1	3.29	4.43	3.86	2.14	1.57	1.57	2.14	2.71
Subject 23	1	2	1	1	1	1	2	1	1	1	5	4	3	1	3	4	5	3
Subject 24	3	4.33	3.67	3.67	5	3	3	4.33	3	3.67	4.33	1.67	1	4.33	3	4.33	1.67	3

4. Experimental results and discussion

The performance was evaluated by assessing how many people would be offered their 'favourite' HRTF datasets by using three different HRTF selection methods. The HRTF datasets with good listening scores in the observed ranking are defined as the 'favourite' HRTF datasets. In the following, we present the experimental results achieved when the HRTF is predicted based on three different methods: the method minimising the total error (Zotkin *et al.*, 2004) as given in Eq. (2), the single-hidden-layer neural network structure as shown in Fig. 10, and the double-hidden-layer neural network structure as shown in Fig. 9.

In Fig. 11, the x-axis indicates the number of top ranked HRTF datasets that are selected and the y-axis indicates the proportion of subjects presented with at least one of their favourite HRTF datasets within the selection. Taking Subject 1's predicted ranking as shown in Table 5 as an example, when the number of top ranked HRTF datasets is three, HRTF 14, 16 and 4 are selected. Because HRTF 16 is actually the fittest dataset for Subject 1 (see the observed ranking in Table 5), we can say that the prediction successfully includes the favourite HRTF dataset if three top ranked HRTF datasets are selected. If we look into another example ranking as shown in Table 6, the first three top predicted HRTF datasets (HRTF 4, 16, and 18) do not successfully include the favourite HRTF dataset (HRTF 17) of Subject 2.

If the favourite HRTFs are defined to be the top ranked HRTF dataset(s) in the observation, HRTF 16

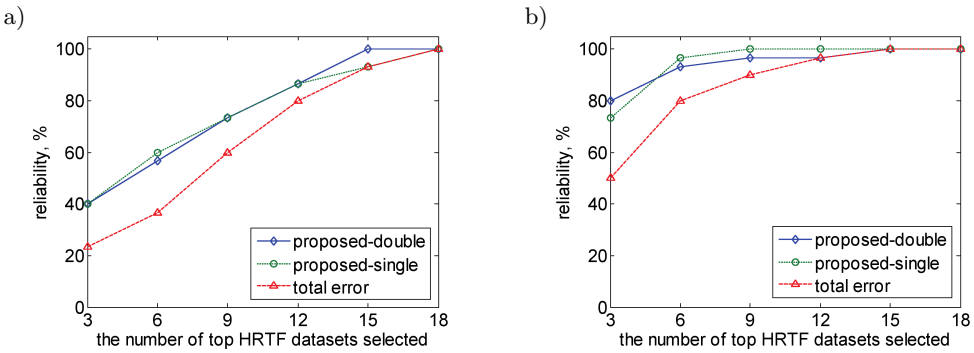

Fig. 11. Performance of different HRTF selection methods when using a given number of top predicted HRTF datasets: a) at least one of the HRTF datasets with first best score was found; b) at least one of the HRTF datasets with first best or second best scores were found.

Table 5. Example ranking of Subject 1.

Predicted HRTF ranking for Subject 1																	
Best score → Worst score																	
HRTF 14	HRTF 16	HRTF 4	HRTF 3	HRTF 17	HRTF 15	HRTF 6	HRTF 18	HRTF 11	HRTF 12	HRTF 5	HRTF 7	HRTF 9	HRTF 8	HRTF 13	HRTF 1	HRTF 2	HRTF 10
Observed HRTF ranking for Subject 1																	
Best score → Worst score																	
HRTF 16	HRTF 18	HRTF 17	HRTF 7	HRTF 6	HRTF 5	HRTF 11	HRTF 8	HRTF 4	HRTF 3	HRTF 1	HRTF 14	HRTF 13	HRTF 12	HRTF 10	HRTF 2	HRTF 15	HRTF 9

Table 6. Example ranking of Subject 2.

Predicted HRTF ranking for Subject 2																	
Best score →																	Worst score
HRTF 4	HRTF 16	HRTF 18	HRTF 7	HRTF 10	HRTF 12	HRTF 14	HRTF 6	HRTF 11	HRTF 15	HRTF 13	HRTF 8	HRTF 9	HRTF 17	HRTF 5	HRTF 3	HRTF 1	HRTF 2
Observed HRTF ranking for Subject 2																	
Best score →																	Worst score
HRTF 17	HRTF 18	HRTF 9	HRTF 7	HRTF 16	HRTF 15	HRTF 14	HRTF 11	HRTF 6	HRTF 3	HRTF 13	HRTF 10	HRTF 8	HRTF 5	HRTF 4	HRTF 2	HRTF 1	HRTF 12

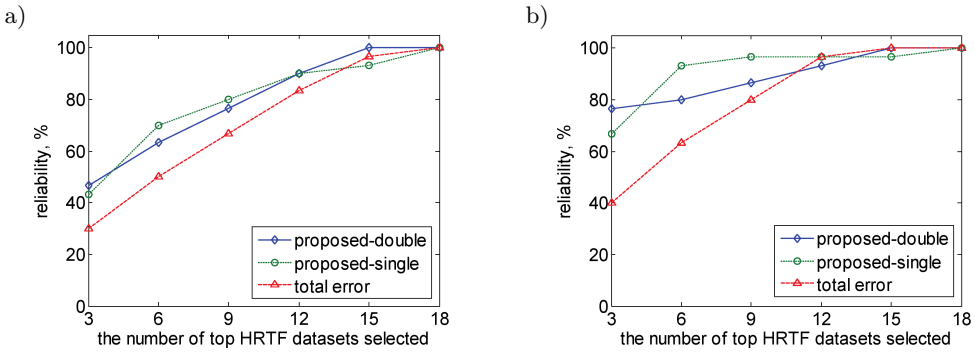

Fig. 12. Performance of different HRTF selection methods when using a given number of top predicted HRTF datasets: a) at least one of the HRTF datasets graded of 4.5 or above was found; b) at least one of the HRTF datasets graded of 4 or above was found.

is the only favourite HRTF dataset for Subject 1. Both HRTF 15 and 18 are called the favourite HRTF datasets for Subject 3, because they owned the equal best listening score (see Table 4). The results in Fig. 11a show that when using the first three top predicted HRTF datasets out of 18, 40% of listeners would be presented with their favourite HRTF datasets when using either of the proposed neural network methods, compared with only 23.3% of listeners when using the total error (Zotkin *et al.*, 2004). If the top and second best ranked HRTF datasets are defined as the favourite HRTF datasets (for example, through looking into Table 4, HRTF 17, 18, 9, and 7 are called the favourite HRTF datasets for Subject 2), the likelihoods that one or more of them are included in the set of selected HRTF datasets are shown in Fig. 11b. The results show that using the first three top predicted HRTF datasets out of 18, 80% of listeners are expected to find one of their favourite HRTF datasets when using the double-hidden-layer neural network. When the number of selected HRTF datasets reaches nine (taking Table 5 as an example, HRTF 16, 18, 17, 7, 6, 5, 11, 8, and 4 are selected for Subject 1; taking Table 6 as an example, HRTF 4, 16, 18, 7, 10, 12, 14, 6, and 11 are selected for Subject 2), the single-hidden-layer neural network is effective for 100% of population. When using the total error method (Zotkin *et al.*, 2004), 15 HRTF datasets would be needed to be selected to satisfy all of the listeners.

We also demonstrate the performance of HRTF selection methods when redefining the set of favourite HRTF datasets. If the favourite HRTF datasets are defined to mean those having the balanced score 4.5 or above (for example, HRTF 5, 6, and 8 are Subject 5's favourite HRTF datasets in Table 4), the achieved performance when at least one favourite dataset is included in the set of selected HRTF datasets is shown in Fig. 12a. If we select the nine top ranked predicted HRTF datasets or fewer, both neural network methods satisfy 10% more participants than the total error method does. If the threshold defining a favourite dataset is reduced to 4 (in this case, HRTF 1, 5, 6, 7, 8, and 9 are Subject 5's favourite HRTF datasets in Table 4), Fig. 12b shows the percentages of population presented with at least one favourite HRTF dataset. The neural network-based methods still outperform the total error-based selection, presenting consistently good reliability. In Fig. 11 and Fig. 12, the performances of the single-hidden-layer structure are similar to those of the double-hidden-layer structure in most cases. Those figures also demonstrate the linear combination of mean squared errors (MSEs) between pinna parameters are not enough when we predict favourite HRTFs.

5. Conclusion

Although an individual's HRTF varies in a complex way with morphological features of human body, we have demonstrated that neural networks can be employed to select the best fit HRTF dataset for a subject

from existing databases. The experimental outcomes also indicate that the previous method based on the sum of the squared error between pinna parameters is not sufficient to accurately predict favourite HRTF datasets.

The anthropometric parameters we selected are from a head model and a pinna model, because these two models cover three important localisation cues: ILD, ITD, and pinna filtering (Collins, 2013). In our future work, the number of features will increase. That is, we plan to extend the parameter set to also include shoulder, torso, and knee reflections, as well as extend the network complexity. It is hoped that this could improve the HRTF ranking prediction further. Moreover, fuzzy rules will be applied to the neural network architecture. This is because one of the criticisms of artificial neural networks is the hardly-justified relationship between inputs and outputs (Dave, Dutta, 2014). There have been several fuzzy rule-based systems proposed to interpret encoded information (Ideri *et al.*, 2004; Benitez *et al.*, 1997; Jang, Sun, 1992). When neural networks are not seen as black boxes, the anthropometric parameters dominating most of HRTFs will be revealed.

The five-grade scale will be replaced by the two-alternative forced choice (2AFC) method, which is a ubiquitous choice for measuring detection or discrimination thresholds (Fechner, 1860) and has been commonly used to test speed and accuracy of choices between two alternatives given a timed interval. By restricting a subject's response to a binary decision, the 2AFC will allow subjects to perform a simpler decision task than the scaling methods in this paper.

Acknowledgments

The authors would like to thank anonymous reviewers for helpful suggestions. The authors would also like to thank all the participants of this study. They volunteered to join the experiments, and spent about 30 minutes to complete the measurements and the listening tests. This study has been in part supported by grant MOST 105-2218-E-305-002 from the Ministry of Science and Technology, Taiwan.

References

1. Algazi V.R., Duda R.O., Thompson D.M., Avendano C. (2001), *The CIPIC HRTF database*, Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Electro-Acoustics, pp. 99–102.

2. Batteau D.W. (1967), *The role of the pinna in human localisation*, Royal Society London, **168**, B, 158–180.

3. Benitez J.M., Castro J.L., Requena I. (1997), *Are artificial neural networks black boxes*, IEEE Transactions on Neural Networks, **8**, 5, 1156–1164.

4. Brown C.P., Duda R.O. (1997), *An efficient HRTF model for 3-D sound*, Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 19–22.

5. Brown C.P., Duda R.O. (1998), *A structural model for binaural sound synthesis*, *Virtual sound rendering in a stereophonic loudspeaker setup*, IEEE Transactions on Audio, Speech, and Language Processing, **6**, 5, 476–488.

6. Choi T., Park Y., Youn D., Lee S. (2011), *Virtual sound rendering in a stereophonic loudspeaker setup*, IEEE Transactions on Audio, Speech, and Language Processing, **19**, 7, 1962 –1974.

7. Chun C.J., Kim H.K., Choi S.H., Jang S.J., Lee S.P. (2011), *Sound source elevation using spectral notch filtering and directional band boosting in stereo loudspeaker reproduction*, IEEE Transactions on Consumer Electronics, **57**, 4, 1915–1920.

8. Collins T. (2013), *Binaural ambisonic decoding with enhanced lateral localization*, Proceedings of Audio Engineering Society 134th Convention.

9. Dave V.S., Dutta K. (2014), *Neural network based models for software effort estimation: a review*, Artificial Intelligence Review, **42**, 2, 295–307.

10. Fechner G.T. (1860), *Elements of psychophysics*, Holt Rinehart & Winston, New York.

11. Gupta N., Barreto A., Joshi M., Aguedelo J. (2010), *HRTF database at FIU DSP lab*, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 169–172.

12. Gupta N., Barreto A., Ordonez C. (2002), *Spectral modification of head-related transfer functions for improved virtual sound spatialization*, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1953–1956.

13. Hagan M.T., Demuth H.B., Beale M. (2002), *Neural Network Design*, CITIC Publishing House, Beijing.

14. Ideri A., Abran A., Mbarki S. (2004), *Validating and understanding software cost estimation models based on neural networks*, Proceedings of IEEE International Conference on Information and Communication Technologies, pp. 433–434.

15. Ircam (2002), *Listen HRTF database*, http://recherche.ircam.fr/equipes/salles/listen/.

16. Jang J.-S.R, Sun C.T. (1993), *Functional equivalence between radial basis function networks and fuzzy inference systems*, IEEE Transactions on Neural Networks, **4**, 1, 156–159.

17. Masterson C., Kearney G., Gorzel M., Boland F.M. (2012), *HRIR order reduction using approximate factorization*, IEEE Transactions on Audio, Speech, and Language Processing, **20**, 6, 1808–1817.

18. Pett M.A. (1997), *Nonparametric statistics for health care research: Statistics for small samples and unusual distributions*, Sage Publications, Thousand Oaks, CA.

19. Ranjan R., Gan W.-S. (2015), *Natural listening over headphones in augmented reality using adaptive filtering techniques*, IEEE/ACM Trans. Audio, Speech and Language Processing, **23**, 11, 1988–2002.

20. Salkind N.J. (2004), *Statistics for people who (think they) hate statistics*, Sage Publications, Thousand Oaks, CA.

21. Shabtai N.R., Rafaely B. (2014), *Generalized spherical array beamforming for binaural speech reproduction*, IEEE/ACM Transactions on Audio, Speech and Language Processing, **22**, 1, 238–247.

22. Tan C.-J., Gan W.-S. (1998), *User-defined spectral manipulation of HRTF for improved localisation in 3D sound systems*, Electronics Letters, **34**, 25, 2387–2389.

23. Watkins A.J. (1978), *Psychoacoustical aspects of synthesized vertical locale cues*, Journal of Acoustical Society of America, **63**, 4, 1152–1165.

24. Wythoff B.J. (1993), *Backpropagation neural networks: a tutorial*, Chemometrics and Intelligent Laboratory Systems, **18**, 115–155.

25. Yao S.-N., Chen L.J. (2013), *HRTF Adjustments with audio quality assessments*, Archives of Acoustics, **38**, 1, 55–62.

26. Zhang M., Tan K.-C., Er M.H. (1998), *A refined algorithm of 3-D sound synthesis*, Proceedings of IEEE International Conference on Signal Processing Proceedings, pp. 1408–1411.

27. Zotkin D.N., Duraiswami R., Davis L.S. (2004), *Rendering localized spatial audio in a virtual auditory space*, IEEE Transactions on Multimedia, **6**, 4, 553–564.